A NEW ESTIMATION FOR THE NUMBER OF UNIQUE POPULATION ELEMENTS BASED ON THE OBSERVED SAMPLE[1]

Philip M. Steel, Bureau of the Census, Washington, D.C. 20233

Keywords: Confidentiality, Unique, Equivalence Classes

I. Introduction.

This paper describes research into a new estimate for the number of uniques in a population based on the information provided by an observed sample. That estimate forms the basis of an estimate for the proportion of uniques in the sample which are unique in the population. The estimate is produced by a fitting algorithm, and much of the work so far has been directed toward producing a good fit. The fitting algorithm is described in detail and results are presented for 25 samples froma known population. While the setting is limited, the results are encouraging.

II. Motivation for problem

One approach to formulating the disclosure risk for a set of data is to measure its diversity, particularly with respect to the set of variables that one suspects are either observable or exist in some more public data set. Statistical agencies are particularly interested in evaluating disclosure risk since much of the data they publish is collected under some pledge of confidentiality. Willenborg and Waal [1996] give a thorough treatment of evaluating disclosure risk.

Before exploring the idea of risk on a set of sample data, lets begin by considering the parent population. If only one individual in the population has a particular combination of values for these variables then the combination of values, or key, is unique and it is likely that the identity of the individual may be deduced. One can classify all the records of the population by their key. If a record is in a class of size one, it is unique, hence at risk of disclosure. If it is in a class with two other records, it is indistinguishable from them with respect to the key and the associated risk is much lower. So a count of unique records is a start in determining the overall risk of the data set. For a sample we not only need to observe the distribution of uniques but also determine what percentage of the unique records observed in a sample are also unique in the parent population, since those that are not unique in the population have lower risk.

III. Background of problem

A strategy for estimating the proportion of uniques uses subsampling, mimicking the original sampling fraction. The number of uniques in the subsample that are unique in the sample is taken as a proxy for the relation of the sample uniques to population uniques. This estimate works reasonably well when the sampling fraction is large. Note that this technique has clear limitations, recursive sampling of any population eventually gives a population consisting only of uniques, so the action of sampling alters the uniqueness distribution.

Another estimate [Zayatz 1991], related to the method presented here, examines the equivalence classes in the sample. The equivalence classes of size 1 are the unique keys. The estimate calculates the probability that an equivalence class of a given size in the population will be represented by an equivalence class of size 1 in the sample and assumes the equivalence class structure of the population is the same as the sample's. The estimate presented here is similar, except we estimate the equivalence class structure of the population by means of a fitting algorithm.

Finding a good estimate for the number of uniques has been a subject of research in the disclosure community for the last 10 years. For a Bayesian approach see [Samuels]. A related problem, estimating the total number of classes has been kicking around for the last 50 years. For recent work on this problem see [Haas Stokes]

IV. Approach

The classification of records by key variables gives one access to a distribution of particular interest: the complete count of classes by size. That is, we will look at $Y = (Y_1, Y_{2,} ..., Y_m)$ where

x_j are individual records and
x_j belong to the same class, $\{x_j\}$,if they have the same key

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This reports is released to inform interested parties of research and to encourage discussion.

$$Y_i=\#\{\text{classes } \{x_j\} \text{ s.t. } \#\{x_j\}=i\},$$

so that Y_1 is the count of unique records, Y_2 the number of different keys which each have 2 records in the data set etc. We will use Y^P to denote the distribution of the population, Y^E to denote the distribution of our estimate of the population, and Y^S to denote the distribution of the sample. Note that $\Sigma(i*Y^P_i)=N$.

What is the relationship between the distribution of the population and the sample? In the example we look at, based on census data, the sampling fraction is taken to be 1/6. For simplicity, we will assume Bernoulli sampling. What is the expected number of unique sample elements? If a record is unique in the population it will be unique in the sample, provided it is selected. That is, we expect 1/6 (p) of the population uniques to wind up in the sample and provide one component of the uniques observed there. Consider a pair of records that comprise a class of two. The probability that they contribute to the class of uniques in the sample is 2pq, ie one is selected and the other is not. Hence the expected contribution from all of the classes of size 2 is $2pq*Y^P_2$. So the expected number of uniques in the sample is

$$p*Y^P_1+2pq*Y^P_2+3pq^2*Y^P_3+ \ldots=\Sigma\, b(p,1,i)Y^P_i$$

where b() are binomial probabilities. More generally

$$(1) \qquad M*Y^P=E[Y^P]$$

where M is an upper triangular matrix with $M(i,j)=b(p,j,i)$ and $E[Y^P]$ is the vector whose entries are the expected number of classes of size i in the sample, given P.

Our objective is to approximate a solution for

$$(2) \qquad M*Y=Y^S$$

where Y^S_i is the distribution of number of equivalence classes of size i observed in the sample.

M is upper triangular hence invertable, unfortunately it is rather large and the determinant is close to 0 so that finding an approximate solution to equation 2 is nontrivial. We will present an iterative fitting algorithm that gives an approximation, Y^E, in the sense that $d(M*Y^E,Y^S)$ is "small" (relative to $\Sigma\, Y^S_i**2$) where $d=\Sigma(Y_i-Y^S_i)**2$.

V. Assumptions.

In order to bring the computational aspect of the problem to a reasonable level, we make some mild assumptions, some of which are specific to our data. Our population consists of records drawn from the 1980 decennial census. The population was initially described in [Zayatz 1991] and was researched in connection with PUMS (Public Use Microdata Samples) disclosure control; its class structure is outlined in table 2. We have already assumed Bernoulli sampling to derive M. We restrict the estimates of the population class size distribution to descending distributions s.t. $\Sigma(i*Y^P_i)=N$. One consequence of the descending assumption in that no estimate which is close to (in the sense above) the true population, can be close in the tails. A typical population will have a key, or a small group of keys, that is significantly more frequent than all others. In our data there are no keys that form any class of sizes 142 through 297, but there is a key that occurs 298 times. The long tail is fairly representative of the distribution in most populations. Barring extremely organized correlation, this is the key that exhibits the most frequent values of all variables that make up the key (eg a household that owns the home, with 2 nonhispanic white adults and 2 nonhispanic white children etc). No descending distribution can fit such a tail, but our concern is to estimate Y^P_1 at the head of the distribution.

We have truncated the M matrix to 100x100. Bounding the tail(s) of the matrix in general is a problem for which we lack a solution. For our known population $d(M_{100}*Y^P-M_{300}*Y^P)<1$, where d is the sum of squared differences between coordinates, and M_{100} is 0 filled to match the size of the larger matrix. The short tail of the estimate tends to further diminish the effect of truncation in the application.

VI. Defining a neighborhood of Y

It is difficult, or at least computationally intensive, to construct a grid of descending distributions s.t. $\Sigma(i*Y_i)=N$. The descending assumption suggests trying a gradient search. That is, by examining nearby distributions, determine which image under M best fits the given sample and iterate.

Recall that Y_i is the number of classes of size i. If Y and Y' differ only in that $Y_5-Y'_5=1$ then the population represented by Y has 5 more units than the population represented by Y'. To preserve additivity and in some sense cover the d-neighborhood of Y, we define an (s,t,n)-neighbor of Y to be:

$$Y_i(s,t,n)=Y_i \text{ for } i\neq s,t$$

$$Y_s(s,t,n)=Y_s+\frac{n}{s}$$

$$Y_t(s,t,n)=Y_t- \frac{n}{t}$$

That is, we will change n units of the population from class s to class t. The resulting distribution still preserves additivity to N. The motivation for a scale parameter is computational: we wish to achieve a coarse fit of the image of the estimate to the given sample before fitting with smaller increments.

VII The fitting algorithm

The algorithm begins from a fixed point, the distribution where every element of the population is unique. In the first iteration only the move of $\frac{n_0}{1}$ uniques to $\frac{n_0}{2}$ doubletons is considered. If that image is closer to the sample, then it becomes the central point and it's neighbors are considered in the next iteration. A neighbor, $Y^{E_i}(s,t,n)$,is adopted only if it's image is closer than the image of Y^{E_i} <u>and</u> it's image is closer to Y^S than any other neighbor. If no neighbor yields an improvement then the size of the increment, n, is decreased. The process is stopped when the increment falls below n_{last}.

More precisely:

Let $Y_1^{E0} = N$ and $Y_j^{E0} = 0$ for j…1 and n_0 be large. Then define

$$Y^{E_{i+1}} = Y^{E_i}(x,y,n_i)$$

where (x,y,n_i) is descending, positive and satisfies

$$d(M*Y^{E_i}(x,y,n_i),Y^S) \le d(M*Y^{E_i}(s,t,n_i),Y^S)$$

œ (s,t,n_i) with $Y^{E_i}(s,t,n)$ descending, positive and

$$d(M*Y^{E_i}(x,y,n_i),Y^S) < d(M*Y^{E_i},Y^S)$$

else $Y^{E_{i+1}} = Y^{E_i}$ and $n_{i+1} = \frac{n_i}{2}$. If $n_i<n_{last}$ then stop.

VII. Results

Twenty five random samples of the parent population were taken and their uniqueness distributions were determined. While not sufficient to draw any solid conclusions about variance and bias, the sets of data have enough variety to work on the efficiency of the algorithm and give an indication of the effectiveness of the estimate. The algorithm described above was applied to each, with n_0=2048 and n_{last}=8. Results seem indifferent to the selection of n_0, but I got a slightly better estimate of the proportion of uniques with n_{last}=32 despite a weaker fit. The results for n_{last}=8 are displayed in tables 2 and 3, the later focusing on convergence. In general, successive iterations take longer and longer to complete: 300 iterations complete in roughly 4 minutes, the 400 in about 10 minutes (On a Pentium II PC, using SAS).

Table 2 show various estimates of the fraction of sample uniques that are unique in the population. The actual fraction is given together with the estimate derived by subsampling, the class based estimator in [Zayatz] and the fitting estimate for the same 25 samples. The table is sorted by the relative difference of the fitting estimate and the actual value. There is a more or less even split of underestimation and overestimation suggesting that variance is more of a problem than bias. In contrast the subsampling estimate and the class based estimate are very regular, but shows considerable bias.

Table 3 is sorted by the number of iterations it took to complete. It shows some variation in fitting the sample distribution and in estimating Y_1^P . The fit of the first three coordinates is given in the table, as well as the total distance from the image of the estimate to the sample. As a gross measure of the regularity of the sample we provide its distance to the expected distribution $M*Y^P$. The most sobering result was that Y_1^S , Y_2^S and Y_3^S were fit more or less exactly for all of the 25 samples, regardless of how good the estimate of Y_1^P was. The estimated population fit the description offered by the sample, even when that estimate was off.

VIII Conclusion and further research.

In order to apply this more generally we need M for simple randomsampling, bounds for the tail of M*Y, and a bound on the error of the estimate itself. The later has two components, first the error associated with treating the sample as an expected value and second the error in approximation.

Convergence might be improved in a number of ways. One could smooth the sample so that it would be easier to fit. A different definition of d may find a more central solution. Relaxing the descending assumption would give a better fit, including the tail.

IX. References

Haas, P. J., Stokes, L. (1998), "Estimating the Number of Classes in a Finite Population", Journal of the American Statistical Association, vol 93,no 444,1475-1487.

Samuels, S.M. (1998), "A Bayesian, Species-sampling-inspired Approach to the Uniques Problem in Microdata Disclosure risk Assessment", Journal of Official Statistics, 14, 373-383.

Willenborg, L., de Waal, T. (1996), Statistical Disclosure Control in Practice, Springer, Lecture Notes in Statistics.

Zayatz, L. (1991), "Estimation of the Number of Unique Population Elements Using a Sample,"Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C.

Table 1 Equivalence Classes in the Data Set

Class Size	Frequency	Class Size	Frequency	Class Size	Frequency	Class Size	Frequency
1	22026	21	18	41	6	61	3
2	2954	22	12	42	5	62	2
3	1090	23	23	43	2	64	4
4	560	24	18	44	1	68	1
5	354	25	15	45	4	69	2
6	223	26	11	46	6	70	2
7	173	27	9	47	3	72	2
8	109	28	7	48	3	75	1
9	106	29	7	49	1	76	1
10	87	30	9	50	2	77	1
11	64	31	8	51	2	78	1
12	53	32	12	52	3	79	1
13	54	33	5	53	3	80	2
14	48	34	7	54	1	86	1
15	26	35	6	55	4	87	1
16	37	36	8	56	1	88	2
17	25	37	7	57	2	101	1
18	14	38	3	58	2	103	1
19	21	39	4	59	1	121	1
20	16	40	3	60	4	141	1
						298	1

Table 2 Estimates of Proportion of Sample Uniques that are Population Uniques
(sort by relative bias of fit estimate)

			Proportion				
			--				
	Distance	Distance				Zayatz	Actual-Fit
Iterations	from	from		Fit	Subsample	Class	------
	$M^{(} Y^E$ to Y^S	Y^S to $M^{(} Y^P$	Actual	Estimate	Estimate	Estimate	Actual
388	217.7	2845.8	0.649	0.732	0.722	0.730	-0.127
207	107.5	601.1	0.662	0.736	0.730	0.731	-0.111
232	145.6	738.3	0.660	0.732	0.740	0.732	-0.109
222	189.1	7228.2	0.664	0.721	0.729	0.734	-0.086
191	128.2	423.9	0.657	0.701	0.729	0.734	-0.067
542	236.0	3726.6	0.644	0.685	0.751	0.739	-0.063
151	108.1	4885.7	0.649	0.690	0.733	0.741	-0.063
456	289.9	1438.7	0.663	0.702	0.739	0.734	-0.059
147	99.9	4620.2	0.661	0.698	0.729	0.731	-0.056
280	154.7	1602.8	0.656	0.689	0.733	0.729	-0.050
169	126.1	3220.9	0.658	0.686	0.731	0.726	-0.042
291	261.6	4214.8	0.656	0.673	0.722	0.726	-0.026
424	166.5	3785.4	0.656	0.650	0.743	0.740	0.008
221	190.3	1758.6	0.657	0.648	0.732	0.736	0.014
191	170.1	700.2	0.658	0.648	0.727	0.729	0.016
522	209.0	418.9	0.658	0.646	0.739	0.732	0.018
621	311.6	884.6	0.661	0.645	0.733	0.731	0.023
748	415.1	2025.7	0.654	0.638	0.741	0.736	0.025

188	61.5	1242.2	0.662	0.644	0.720	0.726	0.027
469	235.4	2562.2	0.652	0.632	0.735	0.733	0.030
215	185.7	1239.6	0.657	0.636	0.722	0.728	0.032
394	318.4	5778.4	0.655	0.628	0.731	0.731	0.041
164	101.2	10254.8	0.661	0.618	0.728	0.736	0.064
365	265.6	10395.6	0.660	0.597	0.719	0.726	0.095
697	338.3	864.2	0.663	0.554	0.740	0.733	0.164

Table 3

Fit and Estimation for 25 Samples from the Population
(sort on fit iterations)

	Distance from image of estimate to sample	Estimate of population uniques	Distance from sample to average sample	The Absolute Fit in the first 3 coordinates		
Iterations	$d\,(M^{(}\,Y^{E}, Y^{S})$	Y^{E}_{1}	$d\,(Y^{S}, M\,(\,Y^{P})$	$M^{(}\,Y^{E}_{1} \& Y^{S}_{1}$	$M^{(}\,Y^{E}_{2} \& Y^{S}_{2}$	$M^{(}\,Y^{E}_{3} \& Y^{S}_{3}$
147	99.9	23092	4620.2	0.1	0.2	1.1
151	108.1	23332	4885.7	0.0	0.3	0.7
164	101.2	21052	10254.8	0.2	0.5	0.5
169	126.1	22740	3220.9	0.1	0.4	0.2
188	61.5	21436	1242.2	0.1	0.1	0.2
191	170.1	12612	700.2	0.3	0.4	0.7
191	128.2	23508	423.9	0.1	0.5	0.7
207	107.5	24612	601.1	0.1	0.3	0.8
215	185.7	21172	1239.6	0.2	0.2	1.1
221	190.3	21564	1758.6	0.4	0.6	1.3
222	189.1	24500	7228.2	0.0	0.2	1.1
232	145.6	24580	738.3	0.2	0.5	0.2
280	154.7	22916	1602.8	0.2	0.6	0.1
291	261.6	22292	4214.8	0.3	0.7	1.2
365	265.6	19628	10395.6	0.3	0.4	2.5
388	217.7	24324	2845.8	0.2	0.3	2.0
394	318.4	21268	5778.4	0.1	0.2	1.0
424	166.5	21884	3785.4	0.3	0.4	2.8
456	289.9	23596	1438.7	0.4	1.1	0.2
469	235.4	21060	2562.2	0.5	0.7	3.5
522	209.0	21652	418.9	0.2	0.1	1.9
542	236.0	23092	3726.5	0.2	0.2	2.2
621	311.6	21668	884.6	0.0	0.1	1.9
697	338.3	18524	864.2	0.4	0.7	3.1
748	415.1	21460	2025.7	0.2	0.3	2.1
	actual	22026				